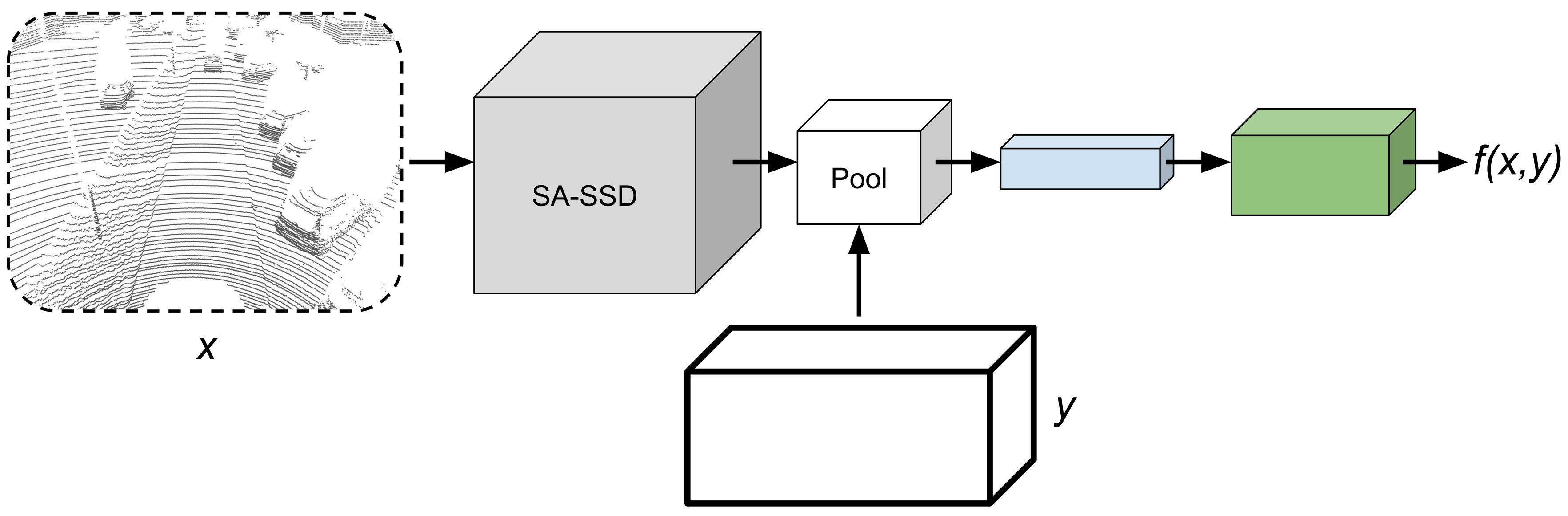
Accurate 3D Object Detection using Energy-Based Models

Fredrik K. Gustafsson[1], Martin Danelljan[2], Thomas B. Schön[1]

[1]Department of Information Technology, Uppsala University, Sweden
[2]Computer Vision Lab, ETH Zürich, Switzerland

Overview

- We extend energy-based regression from 2D to 3D object detection.
- This is achieved by integrating a conditional energy-based model (EBM) $p(y|x;\theta) = e^{f_\theta(x,y)} / \int e^{f_\theta(x,\tilde{y})} d\tilde{y}$ into the state-of-the-art 3D object detector SA-SSD.
- We design a differentiable pooling operator that, given a 3D bounding box y, extracts a feature vector from the SA-SSD output. This feature vector is then processed by fully-connected layers, outputting the scalar energy $f_\theta(x,y) \in \mathbb{R}$.



Energy-Based Regression

Train a neural network $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict a scalar value $f_\theta(x,y) \in \mathbb{R}$, then model the distribution $p(y|x)$ with the conditional EBM $p(y|x;\theta)$:

$$p(y|x;\theta) = \frac{e^{f_\theta(x,y)}}{Z(x,\theta)}, \quad Z(x,\theta) = \int e^{f_\theta(x,\tilde{y})} d\tilde{y}.$$

Energy-Based Regression - Prediction

Predict the most likely target under the model given an input $x^\star$, i.e. $y^\star = \arg\max_y p(y|x^\star;\theta) = \arg\max_y f_\theta(x^\star, y)$. In practice, $y^\star = \arg\max_y f_\theta(x^\star, y)$ is approximated by refining an initial estimate $\hat{y}$ via T steps of gradient ascent,

$$y \leftarrow y + \lambda \nabla_y f_\theta(x^\star, y).$$

Energy-Based Regression - Training using NCE

The neural network $f_\theta(x,y)$ is trained by minimizing the loss $J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} J_i(\theta)$,

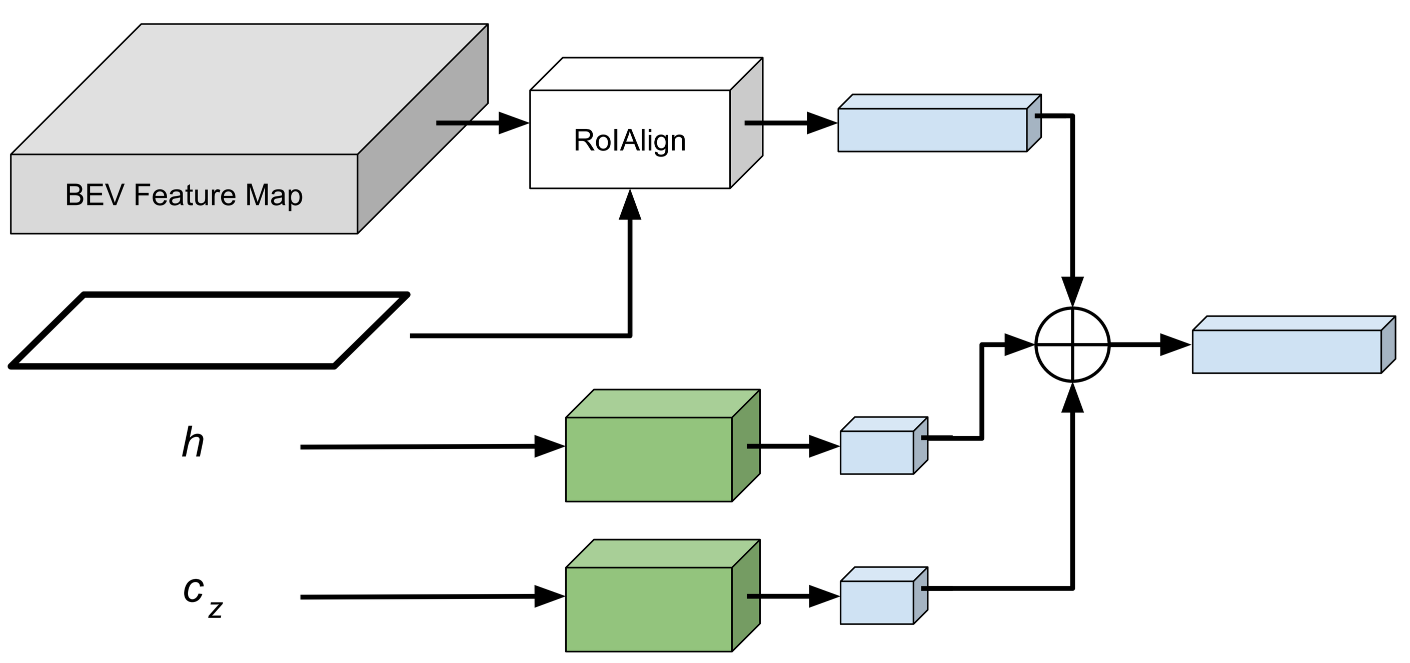
$$J_i(\theta) = \log \frac{\exp\{f_\theta(x_i, y_i^{(0)}) - \log q(y_i^{(0)}|y_i)\}}{\sum_{m=0}^{M} \exp\{f_\theta(x_i, y_i^{(m)}) - \log q(y_i^{(m)}|y_i)\}},$$

where $y_i^{(0)} \triangleq y_i$, and $\{y_i^{(m)}\}_{m=1}^{M}$ are M samples drawn from a noise distribution $q(y|y_i)$ that depends on the true target y_i, $q(y|y_i) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{N}(y; y_i, \sigma_k^2 I)$.

- Effectively, $J(\theta)$ is the softmax cross-entropy loss for a classification problem with $M+1$ classes (which of the $M+1$ values $\{y_i^{(m)}\}_{m=0}^{M}$ is the true target y_i?).

Differentiable Pooling of 3D Bounding Boxes

- The BEV version y^{BEV} of the 3D bounding box y is pooled with the BEV feature map produced by SA-SSD, extracting a feature vector.
- The z coordinate c_z and height h of the 3D bounding box y are processed by two small fully-connected layers, extracting a feature vector each.
- Finally, all three feature vectors are concatenated.



Results on KITTI

TABLE II

RESULTS ON KITTI VAL IN TERMS OF 3D AND BEV AP.

	3D @ 0.7 Easy	3D @ 0.7 Moderate	3D @ 0.7 Hard	BEV @ 0.7 Easy	BEV @ 0.7 Moderate	BEV @ 0.7 Hard
SA-SSD [24]	93.23	84.30	81.36	-	-	-
CLOCs-PVCas [13]	92.78	85.94	**83.25**	93.48	91.98	89.48
PV-RCNN [3]	92.57	84.83	82.69	95.76	91.11	88.93
SA-SSD	93.14	84.65	81.86	96.56	92.84	90.36
SA-SSD+EBM	**95.45**	**86.83**	82.23	**96.60**	**92.92**	**90.43**
Rel. Improvement	+2.48%	+2.58%	+0.45%	+0.04%	+0.09%	+0.08%

TABLE III

FURTHER COMPARISON OF OUR PROPOSED DETECTOR AND THE SA-SSD BASELINE ON KITTI VAL.

	3D @ 0.75			3D @ 0.8			3D @ 0.85			3D @ 0.9		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
SA-SSD	84.48	73.91	70.99	60.89	50.08	47.37	24.29	19.58	18.05	2.06	1.58	1.33
SA-SSD+EBM	87.85	74.96	71.95	66.70	54.32	51.36	31.02	23.91	21.95	3.45	2.74	2.26
Rel. Improvement	+3.99%	+1.42%	+1.35%	+9.54%	+8.47%	+8.42%	+27.7%	+22.1%	+21.6%	+67.5%	+73.4%	+69.9%
	BEV @ 0.75			BEV @ 0.8			BEV @ 0.85			BEV @ 0.9		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
SA-SSD	95.41	87.47	84.79	87.12	79.07	74.65	61.53	54.15	50.39	17.48	15.71	14.58
SA-SSD+EBM	95.47	87.54	84.88	88.31	80.06	77.25	68.40	58.62	54.48	26.60	22.03	19.48
Rel. Improvement	+0.06%	+0.08%	+0.11%	+1.37%	+1.25%	+3.48%	+11.2%	+8.25%	+8.12%	+52.2%	+40.2%	+33.6%

Analysis of Inference Speed

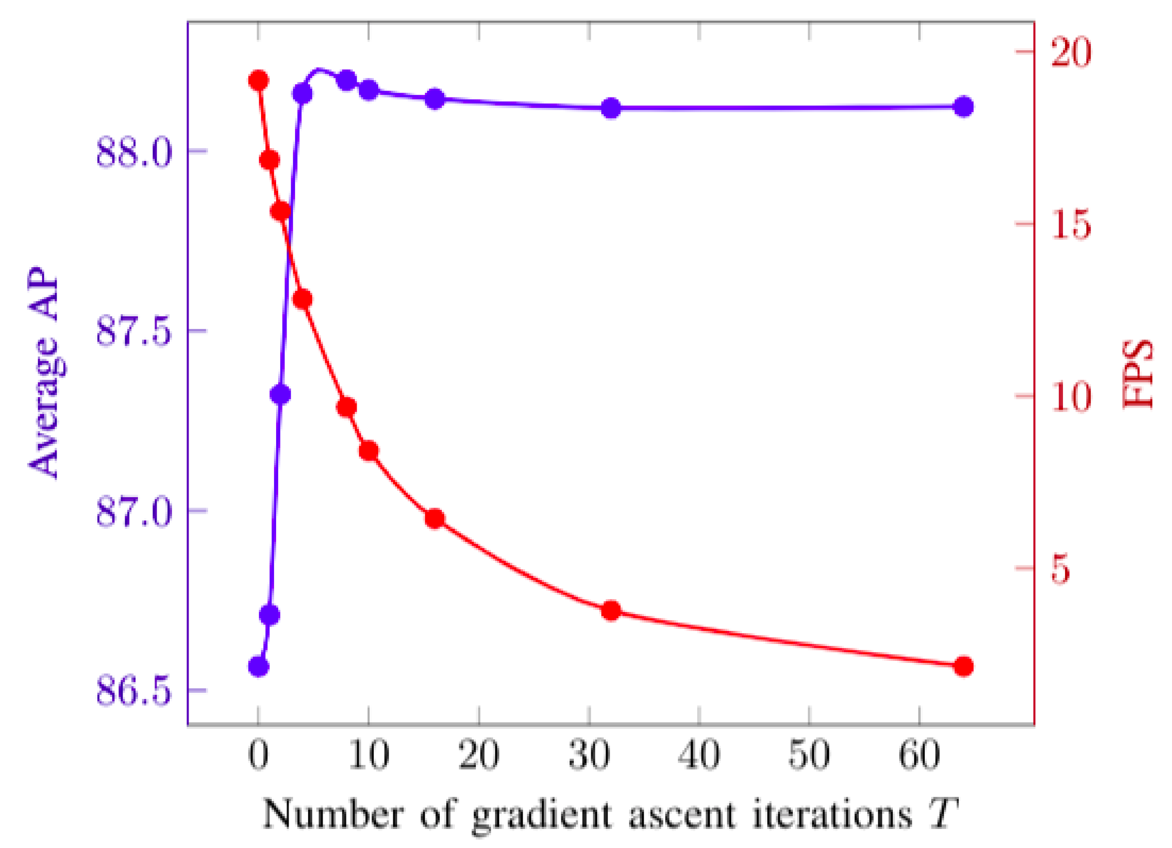


Fig. 5. Impact of the number of gradient ascent iterations T on detector performance (3D AP with 0.7 threshold, averaged over easy, moderate and hard) and detector inference speed (FPS), on KITTI *val*.

Analysis of Learned Distribution

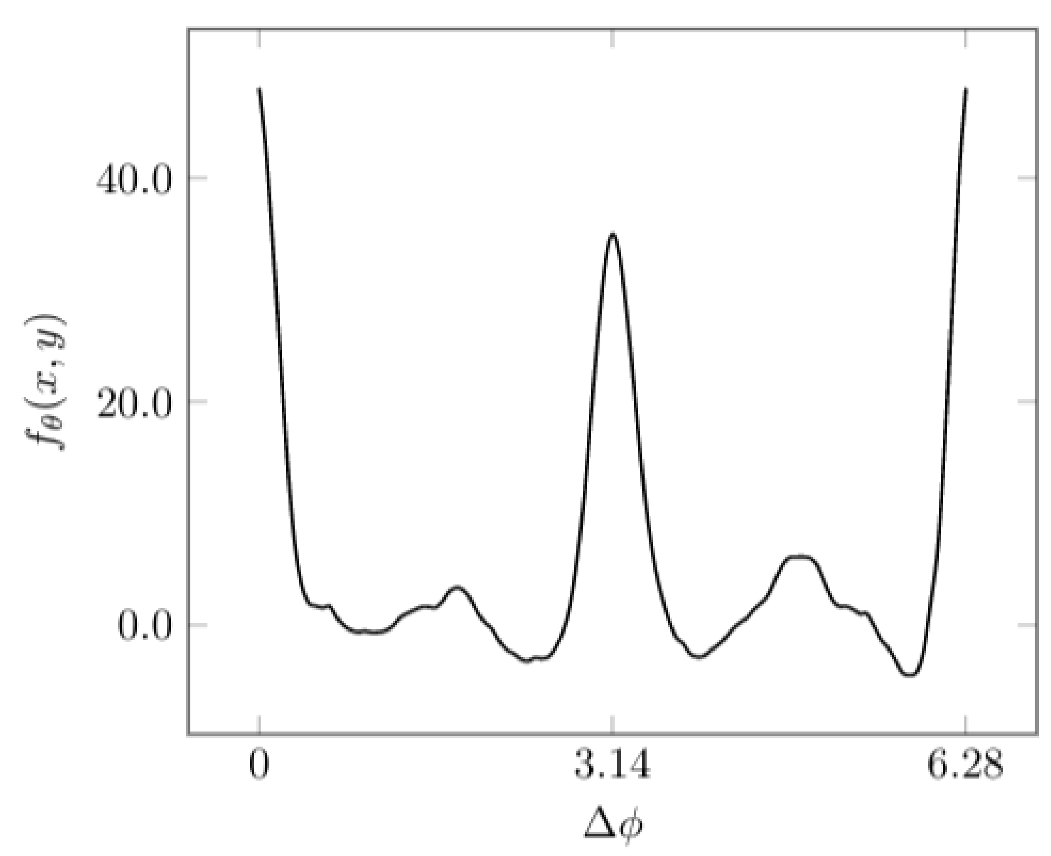


Fig. 6. Visualization of the DNN scalar output $f_\theta(x,y)$ when a predicted 3D bounding box y (6) is rotated $\Delta\phi$ rad, demonstrating that the trained EBM $p(y|x;\theta)$ captures the inherent multi-modality in $p(y|x)$.

Conclusion

- We applied conditional EBMs $p(y|x;\theta)$ to the task of 3D bounding box regression, thus extending the recent energy-based regression approach from 2D to 3D object detection. On the KITTI dataset, our approach consistently outperformed the SA-SSD baseline across all 3DOD metrics, and achieved highly competitive performance also compared to other state-of-the-art methods.
- By demonstrating the potential of energy-based regression for highly accurate 3DOD, we hope that our work will encourage the research community to further explore the application of EBMs $p(y|x;\theta) = e^{f_\theta(x,y)} / \int e^{f_\theta(x,\tilde{y})} d\tilde{y}$ to 3DOD and other important regression tasks.

fredrik.gustafsson@it.uu.se, martin.danelljan@vision.ee.ethz.ch, thomas.schon@it.uu.se